

# Organization Mining Using Online Social Networks

Michael Fire, Rami Puzis, and Yuval Elovici\*

Telekom Innovation Laboratories at Ben-Gurion University of the Negev  
Department of Information Systems Engineering, Ben Gurion University

## Abstract

Mature and developed social networking services are one of the greatest assets of today's organization. However, this great asset is also a non-negligible threat to organization confidentiality. Many details on organizations are exposed on social networking websites by their members along with personal information. In this paper we analyze several commercial organizations by mining data which their employees have exposed on Facebook, LinkedIn, and other publicly available sources. Using a web crawler designed for this purpose, we extract a network of informal social relationships of employees of a given target organization. Our results show that, using centrality analysis and machine learning techniques applied on the structure of the informal relationships network, it is possible to identify leadership roles within the organization. It is also possible to gain valuable non trivial insights on the organizational structure by clustering this network and gathering publicly available information on the employees within each cluster. Organizations willing to conceal their structure, location and specialization of branches, identity of leaders, etc. must enforce strict policies which control the use of social media by their employees.

**Keywords.** Organizational data mining, Social network data mining, Social networks privacy, Organizational social network privacy, Facebook, Machine learning, Leadership roles detection

## 1 Introduction

In recent years, online social networks have grown in scale and variability and today offer individuals the possibility of publicly presenting themselves, exchanging ideas with friends or colleagues, and networking. For example, the Facebook<sup>1</sup> social network has more than 901 million registered users, with new users signing up each month. According to recent statistics published by Facebook, 50% of Facebook users log onto this site on a daily basis, with an average total time of more than 7 hours per month spent online [25] and more than 30 billion pieces of content shared each month (web links, news stories, blog posts, notes, photo albums, etc.) [9]. On the one hand, social networks create new opportunities to develop friendships, share ideas, and conduct business. However, on the other hand, many social network users expose personal third party details about themselves and their social connections via their profile pages [3, 1], in addition to sensitive business information and details about their place of employment.

In this study, we analyze publicly available social network data in order to infer the internal organizational structure of six hi-tech companies of different scales. A similar analysis was performed in the past by Tyler et al. on the HP organization [34]. However, their analysis was based on protected organizational data, i.e., email logs. We show that it is possible to use only publicly available data, from Facebook and LinkedIn, for example, in order to achieve similar results for multiple organizations.

The contributions of this paper are threefold. First, we present a method for uncovering an organization's informal social network topologies based solely on publicly available data. Second, we use the organization's structure to discover hidden leadership roles in the organization and identify communities inside the organization. Lastly, we perform a qualitative analysis of these communities

---

\*Email: {mickyfi,puzis,elovici}@bgu.ac.il

<sup>1</sup><http://www.facebook.com>

and leadership roles and show that it is possible to obtain interesting insights into the organization and the role of each community without any type of access to the organization’s internal data.

## 1.1 Our Approach in a Nutshell

The organization mining methods proposed in this paper were applied to six well-known hi-tech companies of different scales, ranging from small companies with several hundred employees to large scale companies with hundreds of thousands of employees. For each company, the mining process included three major steps. First, we acquired the organization’s informal social network topology from publicly available information, as described in Section 3. As part of this process we collected information about the company’s structure as exposed by the company’s employees on Facebook. The presented method for organizational data mining can help obtain a wide range of an organizations’ social network topologies which were not available to the research community in the past.

Next, we used different centrality measures to detect the hidden leadership roles inside each organization. In Section 4, we highlight the centrality measures with the highest accuracy in pinpointing the leaders. Furthermore, we additionally used Machine-Learning algorithms to classify management roles in each organization.

In the third step, we used a state-of-the-art algorithm in order to cluster the organization’s social network into disjoint communities and cross-referenced the disclosed leaders and communities with the information obtained from LinkedIn (see Section 5). This enabled us to derive the roles of many communities within an organization, providing important insights about the organization. Such insights include, for example, the geographic deployment of the organization, the structure of the different organization’s divisions, the relationships between different divisions and companies that were previously acquired, the research focus of the organization, and more. These details can help us better understand the structure and communication patterns within each organization. If an organization would like to conceal any of these details, it must enforce strict policies which control the use of social media by its employees.

The remainder of this paper is organized as follows. In Section 2, we provide a brief overview of previous relevant studies on social network analysis with a special focus on organizational social network analysis. In Section 3, we describe the methods used in order to obtain the organizational social network structure, and the different organizational datasets obtained. In Section 4, we present methods for identifying an organization’s leadership roles. Next, in Section 5, we describe the methods used to discover the communities roles inside each organization. Lastly, in Section 6, we present our conclusions and offer future research directions.

## 2 Background

In this section, we describe previous work in the fields of online social networks and organizational social networks. We also provide an overview of studies that used different types of informal connections data between an organization’s employees in order to discover the organization’s social network.

### 2.1 Online Social Networks

In recent years, the use of online social networks has grown exponentially. Online social networks, such as Facebook, Twitter<sup>2</sup>, LinkedIn<sup>3</sup>, Flickr<sup>4</sup>, and YouTube<sup>5</sup>, serve millions of users on a daily basis. With this increased use, new privacy concerns have been raised. These concerns result from the fact that online social network users publish personal information about themselves and their work place. In 2007, a study carried out by Dwyer et al. [7] found out that 100% of people who participated in the study used their real name on their Facebook account and 98.6% added photographs of themselves to their Facebook account. Moreover, in 2011, Boshmaf et al. [3] collected and analyzed more than

---

<sup>2</sup><http://www.twitter.com>

<sup>3</sup><http://www.linkedin.com>

<sup>4</sup><http://www.flickr.com>

<sup>5</sup><http://www.youtube.com>

250GB of Facebook users' data and evaluated the amount of personal information exposed by users. They concluded that many Facebook users disclose personal information about themselves, including dates of birth, place of work, email addresses, relationship status, and even phone numbers. By using publicly available data from Facebook and cross-referencing it with other public data sources on the web, such as Google<sup>6</sup> and LinkedIn, one can infer details about a Facebook user, such as work experience and expertise. For example, Pipl<sup>7</sup> and PeekYou<sup>8</sup> are able to search for information about people across different social networks. These *people search engines* aggregate the obtained results and present a fully detailed personal profile.

In this study we use publicly available data from Facebook in order to identify which Facebook users worked for a specific organization. We then cross-reference the users' details with LinkedIn, Google search results, and the company's web-page, in order to reveal the users' positions in the organization.

## 2.2 Organization's Social Networks Structure Analysis

In the past six decades, a considerable amount of research has gone into analyzing and understanding communication patterns between individuals inside organizations. In 1951, Jacobson and Seashore [15] were among the first researchers to study communication patterns among federal agency organization employees. In 1968, Pugh et al. [28] studied five primary dimensions of the organizational structure on 52 different organizations in England. In 1969, Allen and Choen [2] studied technical communication patterns and their influences in two research and development laboratories at MIT. In 1979, Tichy et al. [33] presented a method for analyzing organizations using a network framework which included many network structure properties, such as centrality, clustering, and density. Tichy et al. used their framework to perform a comparative analysis of two organizations with several hundred employees. In 1991, Sparrow [31] presented a method for using social network structure analysis in order to better understand criminal organizations. In 2002, after the tragic events of September 11, 2001, Krebs [19] studied Al-Qaeda's organizational network structure properties and succeeded in identifying the conspiracy leader by using the degree and closeness structural properties of vertices. In 2003, Campbell et al. [5] presented algorithms for expertise identification by using email communication patterns. Their algorithms were evaluated on two different organizations. In our study, we show that expertise, leadership, and the roles of communities can be identified using publicly available data sources even without having access to internal organization data, such as email logs.

In recent years, and with the increasing popularity of online social network usage, many studies have been performed on the usage and benefits of public and internal social networking services to organizations. In 2009, Steinfield et al. [32] studied the connection between social capital and the usage of internal social networking services deployed inside organizations. In the same year, Rooksby et. al. published a detailed report on how online social networks are used in the context of the work place [29]. Comprehensive reviews on organizational social networks can be found in [17, 27, 16].

## 2.3 Discovering an Organization's Social Network from Informal Connections

The work reported in this paper is closely related to the 2004 internal study on the HP organization carried out by Tyler et al. [34]. By analyzing the organization's email corpus, which contained more than one million messages, they discovered the organizational social network topology and identified communities inside the organization. The authors used the betweenness-centrality measure [11] in order to detect leadership roles inside the organization. They also applied a version of the Wilkinson and Huberman algorithm [35] which partitions the organization's social network into communities. They evaluated their results by interviewing several employees about the community they were placed in by the community detection algorithm. Naddaf and Mutyala [36] presented a similar study in 2010. They demonstrated a method for extracting informal social networks formed by the employees of an organization based on the employees' email records. They tested their method on a large public sector

---

<sup>6</sup><http://www.google.com>

<sup>7</sup><http://pipl.com>

<sup>8</sup><http://www.peakyou.com>

client and identified the authority of the employees by using the PageRank measure [26]. Moreover, Naddaf and Mutyala used the Fast Modularity algorithm [6] in order to identify communities in their client’s organizational social network.

### 3 Organization Social Network Crawler

In recent years many different types of web crawlers have been developed in order to collect data from large scale online social networks [21, 3, 12, 10]. Usually, social networks crawlers start from several seed profiles and gradually expand the set of acquired profiles using, for example, Breadth-Search-First (BFS) crawling or other methods, such as Random-Walks [12].

Unfortunately, standard crawling techniques are insufficient for performing data collection which focuses on a specific organization. During the preliminary study performed using BFS crawling, we collected many irrelevant profiles and skipped Facebook users who worked in a target organization. To tackle the problem of targeted acquisition of profiles from online social networks, we developed an organization crawler which optimizes data collection from users associated with a specific group or organization. According to the homophily principle [20], it is more likely that a person has been employed by a certain organization if many of his friends have been employed by the same organization as well.

In order to mine the social network for the profiles of employees from some target organization, our crawler worked according to the algorithm depicted in *Algorithm 1*. The crawl starts from a set of seed profile pages initially identified as belonging to employees of the targeted organization. The initial set of seeds can be obtained using a search engine. These seeds are used to initialize a priority queue (line 2). All seeds have an initial priority of zero. Later, the priority of profile pages in the queue is increased with every friend that is employed by the target organization (lines 11-12).

We proceeded by iteratively processing the next profile page with the highest priority (i.e., likelihood of being an employee of target organization) until no potentially valuable profiles are left in the priority queue (lines 4-5). Every processed profile page is downloaded (line 7) and automatically analyzed. We employed a heuristic in an attempt to discover whether the currently processed profile page belongs to an employee of the target organization. This heuristic matches various keywords associated with the organization to the semi-structured data that appears in the user’s publicly available Facebook profile. For example, in order to identify users from Ben-Gurion University’s Information System Engineering Department, the crawler searches for strings such as “Ben-Gurion Information System Engineering”, “BGU ISE”, or “ISE BGU”. in the collected profile page. In case we do find a match to either one of the keywords, we continue on to the next profile in the priority queue. However, if we find that the dequeued profile page belongs to an individual that had worked in the targeted organization, we collected the list of his Facebook friends (lines 8-9).

Profile pages of Facebook friends that were already processed are ignored (line 10). We increased the priorities of all friends waiting in the priority queue (lines 11-13). Afterwards, we inserted all newly encountered Facebook friends of the currently processed profile into the priority queue, with a priority of one (lines 14-16). This process repeats with the next profile page extracted according to the updated priorities.

The crawler whose pseudo code is described by Algorithm 1 stops when the queue is empty. We will refer to this crawler as Version 1. We also evaluated an optimized version of the organization crawler. This version tracks the number of friends within the targeted organization that each user profile in the priority queue has and the number of organization’s employees discovered during the last iterations. We stop the crawling process if all users in the priority queue have at most one friend in the targeted organization and the last thousand profiles acquired from Facebook did not belong to the organization’s employees. We will refer to the crawler with this stricter stopping condition as Version 2.

---

**ALGORITHM 1:** Organization Social Network Crawler (Version 1)

---

**Input:** A set of seed URLs ( $S$ ) to Facebook profile pages of organization’s employees.  
A set of crawling organization target names,  $N$ .  
**Output:** A set of Facebook profiles and their connections.  
 $Q \leftarrow \text{Priority-Queue}()$   
 $\forall_{URL \in S}, Q.\text{Enqueue}(URL : 1)$   
 $Crawled \leftarrow \emptyset$   
**while** ( $Q \neq \emptyset$ ) **do**  
     $URL \leftarrow Q.\text{Dequeue}()$   
     $Crawled \leftarrow Crawled \cup \{URL\}$   
     $Page \leftarrow \text{DownloadProfilePage}(URL)$   
    **if**  $Page$  contains  $N$  **then**  
         $F\_URLs \leftarrow \text{Extract list of friends from Page}$   
         $F\_URLs \leftarrow F\_URLs - Crawled$   
        **for** ( $F\_URL \in F\_URLs \cap Q$ ) **do**  
            Increase priority ( $Q, F\_URL$ )  
        **end**  
        **for** ( $F\_URL \in (F\_URLs - Q)$ ) **do**  
             $Q.\text{Enqueue}(F\_URL:1)$   
        **end**  
    **end**  
**end**  
**return** Collected pages

---

### 3.1 Collected Organizations Datasets

In order to test the methods of organization data collection reported in Section 3, we used the organization social network crawler to collect publicly available data of six well-known hi-tech companies of different scales. The organization crawling results are depicted in Table 1.

Using the organization social network crawler, we collected data from companies on three different scales: Small (S), Medium (M), and Large (L) scale companies currently employing 500 to 2,000, 4,000 to 20,000, and more than 50,000 employees, respectively. Data on one company of each scale was acquired using each version of the organization crawler. We refer to the three companies targeted by Version 1 and Version 2 of the crawler as S1, M1, L1 and S2, M2, and L2, respectively. In the rest of this section, we describe in detail the properties of each collected organization dataset (See Table 2). We used Cytoscape [30] software to visualize the social networks formed by the employees of each organization. The vertex colors in Figures 1-6 represent the various cluster roles, as will be explained in Section 5. The analysis results of this networks are reported in Sections 4 and 5.

Table 1: Organization Crawling Results

Org.	Crawler Version	#Total Crawled Profiles	#Org. Crawled Profiles	Precision
S1	Version 1	22,992	165	0.7%
S2	Version 2	3,312	320	9.6%
M1	Version 1	11,247	1,429	12.7%
M2	Version 2	7,422	3,862	52%
L1	Version 1	13,505	5,793	42.9%
L2	Version 2	18,810	5,524	29.3%
<b>Total</b>	<b>-</b>	<b>77,288</b>	<b>17,096</b>	<b>22.1%</b>

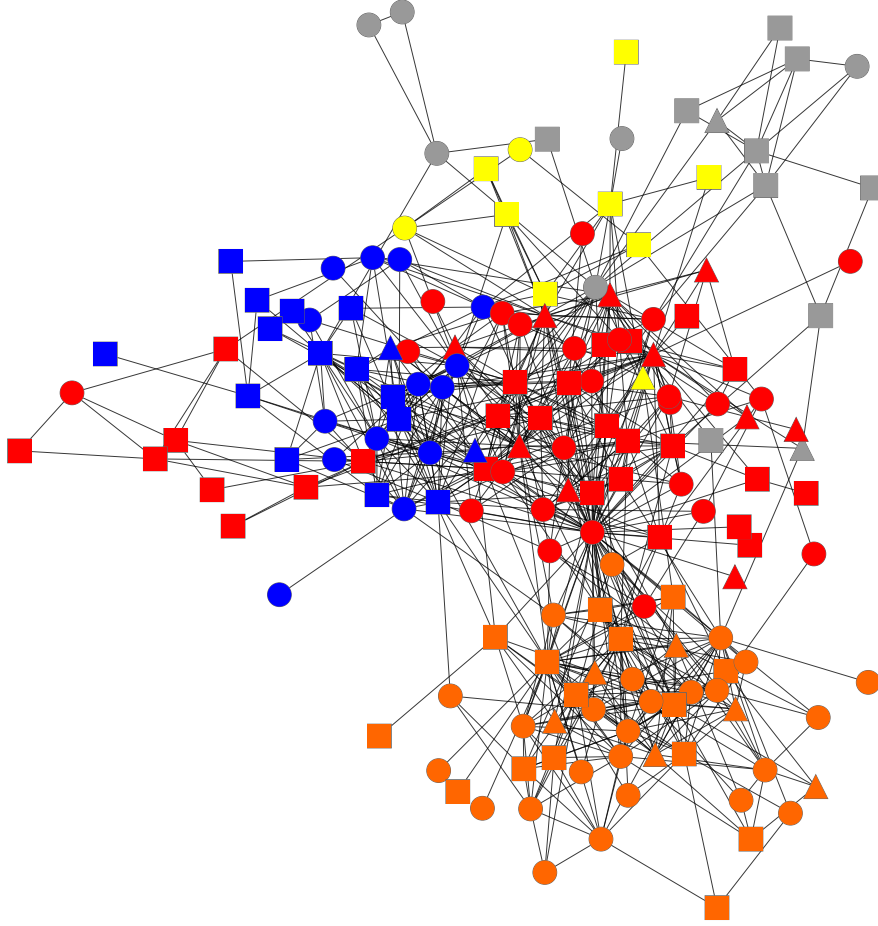


Figure 1: **S1 Company:** *Blue nodes-* R&D and administration groups in Asia, *Red nodes-* mainly hardware verification engineers and chip designers in Asia, *Yellow nodes-* Hardware R&D, *Orange nodes-* acquired startup company, *Gray nodes-* R&D in Asia.

### 3.1.1 International Small Hardware Company (S1)

The S1 company is a publicly held company that specializes in network hardware development. According to the company's web page, the company employs between 500 to 1,000 employees and holds two headquarters located in North America and Asia. We used the organization crawler to identify 726 informal links between 165 Facebook users who, according to their Facebook page, worked in the company. We also collected information on the positions inside the company of 84 employees. Out of 84 employees, we identified 20 in management positions. Most of the discovered company employees held R&D positions, while most of the identified managers were R&D team leaders.

### 3.1.2 International Small Software Company (S2)

The S2 company is an internationally publicly held company that specializes in software development. According to public sources, the company employs between 1,000 to 2,000 employees and holds several headquarters located in North America, Europe, Asia, Australia, and the Middle East. We used our organization crawler to identify 2,369 informal links between 320 Facebook users who, according to their Facebook profiles, worked in the company. We also collected information on the positions of 168 company employees. Out of the 168 employees, we identified 76 in management positions. Many of the company employees held project-manager (PM) positions, while we also identified a similar number of developers, QA, and support employees.

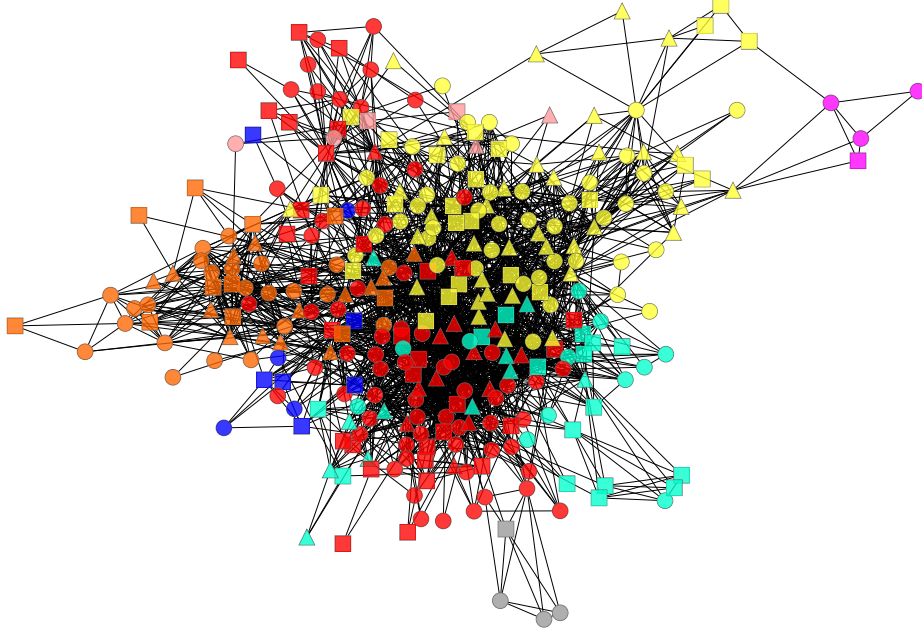


Figure 2: **S2 Company:** *Blue nodes-* IT group in the Middle East, *Red and Orange nodes-* R&D groups in the Middle East, *Purple nodes-* Group in North America, *Yellow nodes-* managers and international project managers, *Cyan nodes-* R&D teams in Australia and the Middle East, *Gray nodes-* European group.

### 3.1.3 International Medium Telecommunication Service Company (M1)

M1 is an international technology company located in North America that specializes in telecommunication services. According to the company's web page, it currently employs between 2,000 and 10,000 employees. We used the organization crawler and identified 32,876 informal links between 1,429 Facebook users who, according to their Facebook profile page, worked in the company. We also collected information on the positions of 461 employees. Out of the 461 employees, 227 hold management positions. A wide range of positions inside the company were identified during the crawl: Senior management positions, sales and marketing positions, PMs, developers, IT engineers, support engineers, technical writers, etc.

### 3.1.4 International Software Provider and Outsourcing Company (M2)

M2 is an international software and outsourcing provider that specializes in telecommunication services and serves customers across the world. According to the company's web page, the company currently employs between 10,000 and 20,000 employees. We used the organization crawler to focus on the company headquarters in South Asia. We stopped the crawling process after identifying 87,324 informal links between 3,862 Facebook users who, according to their Facebook profile page, worked in the company. We also succeeded in collecting information on the positions within the company of 1,511 employees. During the crawl a wide range of positions were identified: Senior management positions, sales and marketing positions, developers, IT, PM, support engineers, technical writers, etc. Out of the 1,511 employees, 230 held management positions.

### 3.1.5 Large Information Technology Corporation (L1)

L1 is an international IT Corporation that provides products and services to customers around the world. According to the company's web page, the company currently employs more than 50,000 employees. We used our organization crawler to collect data on corporation employees in South and North America, Asia, Eastern Europe, and Asia. We stopped the crawling process after identifying

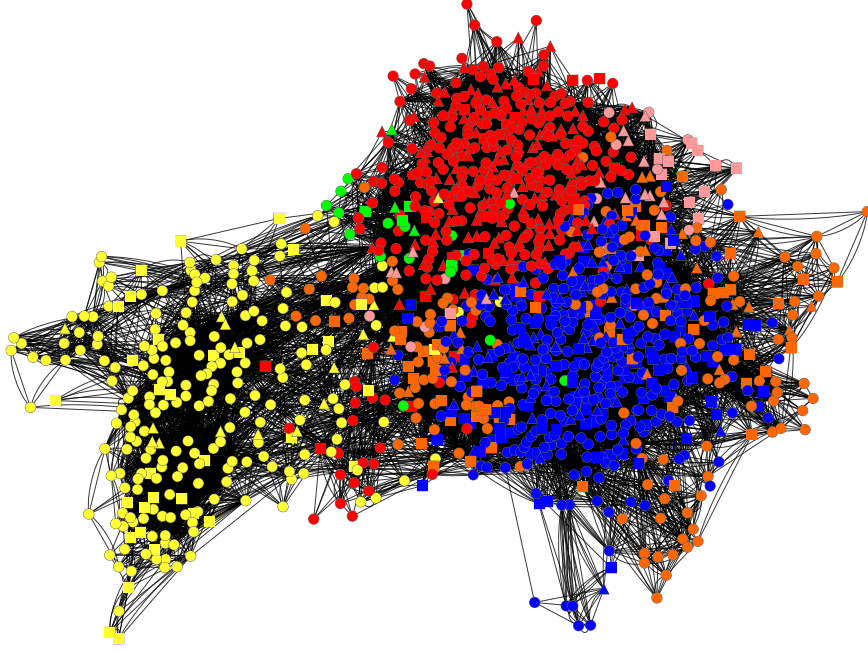


Figure 3: **M1 Company:** *Blue and Orange nodes-* R&D divisions, *Red nodes-* senior management, *Yellow nodes-* International consultants and support engineers, *Green nodes-* North American Headquarter.

45,266 informal links between 5,793 Facebook users who, according to their Facebook profile page, worked in the corporation. We also succeeded in collecting information on the company positions of 1,619 employees. Out of 1,619 employees, we succeeded in identifying 463 holding management positions. A wide range of positions were identified in different parts of the globe: Senior management positions, sales and pricing positions, marketing positions, developers, IT, PM, support engineers, technical writers, etc.

### 3.1.6 Large Technology Corporation (L2)

L2 is an international Technology Corporation that provides hardware and software products to customers, infrastructure, and other services to customers around the world. According to the company's web page, the company currently employs more than 50,000 employees. We used our organization crawler to collect data on corporation employees in South and North America, Asia, Eastern Europe, and Asia. We stopped the crawling process after identifying 94,219 informal links between 5,524 Facebook users who, according to their Facebook profile page, worked in the corporation. We also succeeded in collecting information on the company positions of 1,131 employees. Out of 1,131 employees, we succeeded in identifying 461 management positions. During the crawling, we identified a wide range of positions inside the company: Senior management positions, sales and marketing positions, project managers, developers, IT, support engineers, technical writers, etc.

## 4 Identifying Organization's Leadership Roles

After the organization crawler completes collecting data from the Facebook profiles of employees of the targeted organization, we can analyze the organizational social network created by the "informal" Facebook connections. Let  $G = \langle V, E \rangle$  represent the informal social network, where node  $v \in V$  is a Facebook user who worked in the target organization and  $(u, v) \in E$  represents a Facebook friendship link between two users.



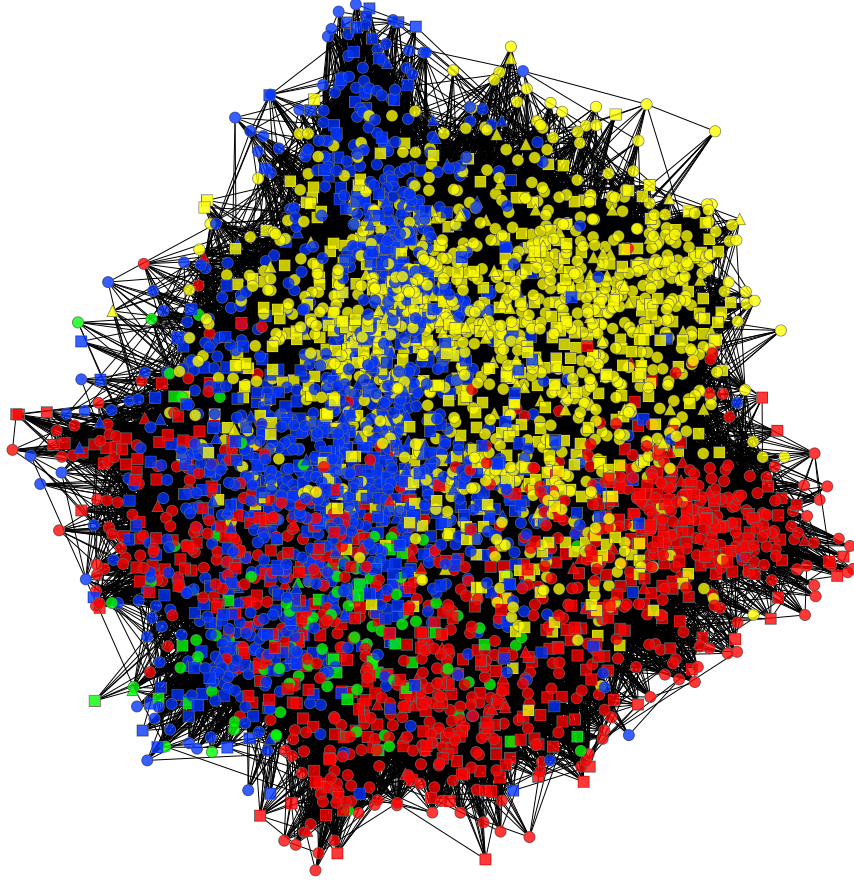


Figure 4: **M2 Company:** *Blue and Green nodes* - R&D and Specific Domain Experts (SDE) connected to North American and Asia employees, *Red Nodes* - R&D and Specific Domain Experts connected to Australia, Europe and North America, *Yellow nodes* - R&D and Specific Domain Experts connected to Africa, North America, and Asia.

In this section we demonstrate that it is possible to pinpoint leadership roles by analyzing solely the structure of the informal social network of organization’s employees. First, for each user  $v \in V$  in the informal social network, we calculated eight centrality measures. Next, for each centrality measure, we examined the top 10 and the top 20 users who received the maximal score.

Using the user’s Facebook profile, we manually classified whether the user held a management position. However, in many cases the user’s profile information was not enough to reveal the user’s positions inside the organization. To overcome this problem, we cross-referenced the user’s personal details with other publicly available online sources, such as LinkedIn and Google search engines. By using these methods, we succeeded, in many cases, to reveal the user’s positions inside the organization. Lastly, we used several Machine-Learning algorithms to build classifiers that can automatically identify management roles inside an organization based on the different centrality measures of the vertices in the informal social network. By using these classifiers, we can recall a wider range of management roles that answer complex centrality measures criteria. Moreover, these types of classification methods can be used to compromise users’ privacy by exposing hidden positions inside the organization. Furthermore, similar methods can assist in revealing different statistics about the organization, and by this compromise the organization’s privacy. For example, using the above methods, we estimated the percent of management positions and the number of employees inside each organization (see Tables 2 and 4). In many privately held organizations, this type of data may be confidential organizational information.

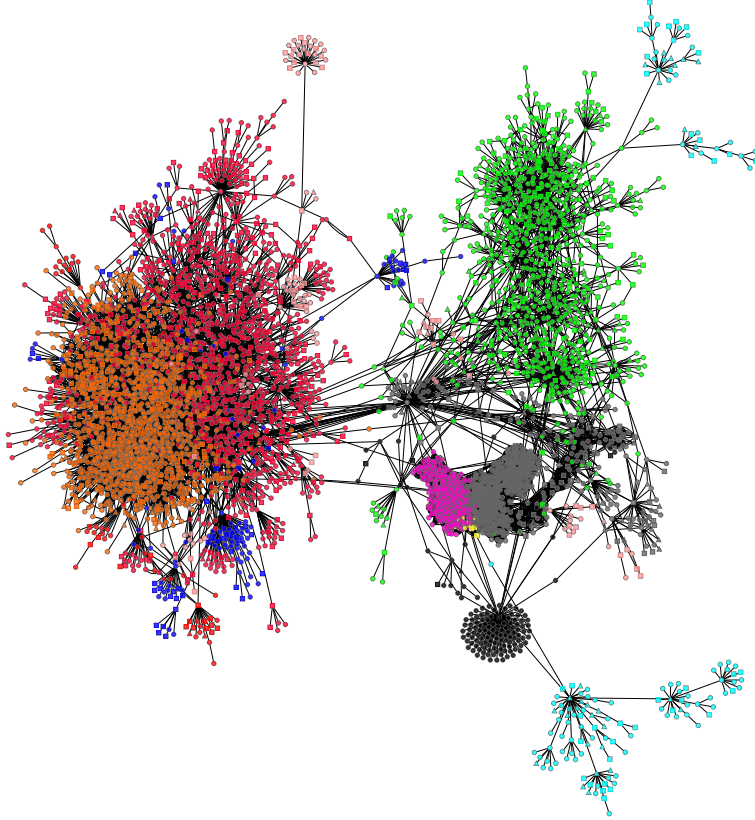


Figure 5: **L1 Corporation:** *Blue nodes*- South American support engineers, *Red nodes*- South American Branch (IT, Support engineers, Analysts, and PMs), *Orange nodes*- South American Branch (Management, Sales, Marketing, Project Managers, Support engineers, and Administration), *Yellow nodes*- Eastern Europe Pricing Analysts, *Purple nodes*- Eastern European (Marketing, Sales and Pricing) consultants and support engineers, *Black nodes*- North American Branch, East Asia - R&D, *Green nodes*- Middle East R&D and North American Headquarters (Management and Sales), *Gray nodes*- European Consultants and Sales and South Asian Analysts.

#### 4.1 Centrality Measures

Using the organization datasets described Section 3.1, we attempted to identify leadership roles within the organization using several centrality measures. For each node in the informal organization social network, we calculated eight centrality measures<sup>9</sup>: Degree centrality (DG), Closeness centrality (CL) (Closeness) [22], Betweenness centrality (BC) [11], Eigvector centrality (EC) [23], HITS (H) [18], PageRank (PR) [26], Communicability centrality (CC) [8], and Load centrality (LC) [24]. We then sorted the crawled organization’s users list according to the different centrality measures.

We manually inspected the top 20 user profiles according to each centrality measure in order to infer their position within the target organization. As a large fraction of Facebook users do not disclose their positions on their profile page, we used other online sources, such as LinkedIn or results returned by Google search engine, in order to manually classify whether a particular employee holds a management position. We will refer to managers who do not report their position on Facebook as concealing their management position. We show here that the location of employees in the informal social network of the organization reveals their management role within the organization with high precision, even though it is not reported on Facebook.

<sup>9</sup>The centrality measures were calculated by using the Networkx [13] Python package.

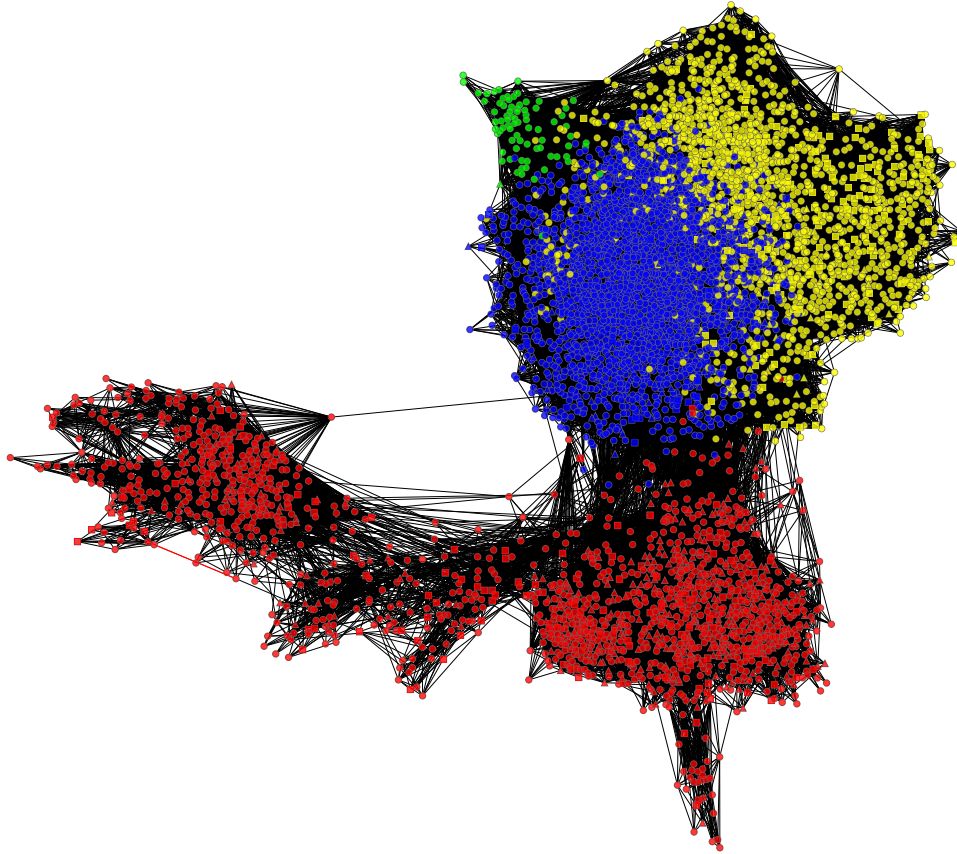


Figure 6: **L2 Corporation:** *Blue nodes*- East Asia Headquarter (management and consultants), *Red nodes*- International Senior Management (Senior Sanagement, Senior researchers), *Yellow nodes*- East Asian Headquarter (R&Ds and consultants), *Green nodes*- The company’s amateurs sport team.

Table 3 presents the leadership identification *precision* at the top 10 and at the top 20 for the various centrality measures. The results indicate that each of the calculated centrality measures can assist in identifying managers inside the organizations. Closeness demonstrated the highest average precision at 20 (0.78), while PageRank received the lowest score (0.7).

Table 4 reports the number of concealed management roles that can be detected by using the closeness measure. Out of 85 managers detected by focusing on the top 20 Facebook users with the highest Closeness centrality within the informal social network of their organization, 40% did not report their positions on Facebook.

According to these results, high centrality within the informal social network of an organization is a good indication of a leadership role within an organization. However, this straightforward general method can only identify management roles of employees with relatively high centrality measures and therefore other management roles with more complex centrality criteria will not be identified using this method. To overcome this problem of identifying management roles with more complex centrality criteria, we used state-of-the art Machine-Learning algorithms to classify management roles in each organization (see Section 4.2).

## 4.2 Machine Learning

In this study we also used state-of-the-art Machine-Learning techniques and constructed classifiers that can identify management positions inside each organization. By using these type of Machine-Learning techniques, we can identify employees with management roles who satisfied more complex centrality criteria. Moreover, using similar methods and techniques can assist in identifying employees

Table 2: Collected Organization Datasets

Org.	Size	Discovered Em- ployees	Links	Employees Disclosing Positions on Facebook
S1	500-1K	165	726	54(32.7%)
S2	1K-2K	320	2,369	104(32.5%)
M1	2K-10K	1,429	32,876	383(26.8%)
M2	10K-20K	3,862	87,324	1,531(39.6%)
L1	50K+	5,793	45,266	1,601(27.6%)
L2	50K+	5,524	94,219	1,131(20.5%)
<b>Total</b>	<b>-</b>	<b>17,093</b>	<b>262,780</b>	<b>4,804(28.1%)</b>

Table 3: Management Positions Percentage Based on Centrality Measures (Precision at 10/20)

Org.	Cat.	DG	CL	BC	H	PR	EC	CC	LC
S1	T-10	0.5	0.4	0.6	0.3	0.5	0.3	0.3	0.6
	T-20	0.35	0.3	0.3	0.3	0.25	0.3	0.3	0.3
S2	T-10	0.8	0.9	0.8	0.9	0.7	0.9	0.9	0.8
	T-20	0.7	0.75	0.75	0.7	0.75	0.7	0.75	0.75
M1	T-10	1	1	0.8	1	1	1	1	0.8
	T-20	1	0.95	0.85	1	0.85	1	1	0.85
M2	T-10	0.83	0.71	0.86	0.83	0.86	0.83	0.83	0.88
	T-20	0.73	0.82	0.69	0.8	0.71	0.8	0.8	0.69
L1	T-10	0.55	0.8	0.8	0.78	0.6	0.78	0.78	0.8
	T-20	0.65	0.75	0.7	0.56	0.65	0.56	0.56	0.7
L2	T-10	1	1	1	1	1	1	1	1
	T-20	0.92	1	1	1	1	1	1	1
<b>Avg.</b>	<b>T-10</b>	<b>0.78</b>	<b>0.8</b>	<b>0.81</b>	<b>0.8</b>	<b>0.78</b>	<b>0.8</b>	<b>0.8</b>	<b>0.81</b>
	<b>T-20</b>	<b>0.725</b>	<b>0.76</b>	<b>0.715</b>	<b>0.73</b>	<b>0.7</b>	<b>0.73</b>	<b>0.735</b>	<b>0.715</b>

with different types of positions, such as Senior management positions and R&D engineers. In order to use the Machine-Learning algorithm, we first needed to create a training set consisting of sufficient training instances. Every training instance represented a collected user in the organization. The target attribute is a binary attribute which indicates whether the user held a management role inside the organization, while the instance features were the different extracted centrality measures. We created a sufficient number of training instances by quickly reviewing the users' data extracted from the users' Facebook profiles. By analyzing the crawled organizations' user Facebook profiles, we discovered that an average of 28.1% of the collected organizations' users inserted at least partial information about their previous and current work experience positions into their Facebook profile page (see Table 2). For each user who inserted their previous work experience into his Facebook profile page, we attempted to determine if the user held a management role inside the organization. In some cases we also did a deeper inspection of the user by cross-referencing the user's work experience with data obtained from other sources, such as LinkedIn. Using this method, we reviewed and classified 4,767 users' profiles. Out of 4,767 manually classified positions, we identified 1,470 users who held management positions (see Table 4). All these profiles were fed into WEKA [14], a popular suite of Machine-Learning software, as training instances. By using WEKA we tested many different Machine-Learning algorithms, such as *OneR* (OR), *K-Nearest-Neighbors* (*IBk*) with  $K \in \{1, 3, 10\}$ , *Naive-Bayes* (NB), *Decision tree* (*J48*), *Logistic* (LG), and *RandomForest* (RF). Lastly, we evaluated each classifier by using the 10-folds cross validation method and calculating the *Accuracy*, *F-measure*, and *AUC* (Area Under the ROC curve) (see Table 5). Next, we used T-tests with a significance of 0.05 to compare

Table 4: Organization’s Hidden Management Positions

Org.	Classified Employees Positions	Classified Management Positions	Closeness T20 Management Positions	Hidden T20 Management Positions
S1	84	20 (23.8%)	6	5 (83.33%)
S2	168	76 (45.2%)	15	4 (26.66%)
M1	461	227 (49.2%)	19	10 (47.4%)
M2	1,511	223 (14.76%)	14	2 (14.3%)
L1	1,619	463 (28.6%)	15	3 (20%)
L2	924	461 (49.9%)	16	10 (62.5%)
<b>Total</b>	<b>4,767</b>	<b>1,470 (30.8%)</b>	<b>85</b>	<b>34 (40%)</b>

the different classifiers. According to T-test results, for all organizations except S1, all the classifiers returned better accuracy results than the naive ZeroR (ZR) classifier. Moreover, in most cases, the simple OneR classifier is sufficient enough to obtain a near maximum accuracy. However, better AUC and F-Measure results were obtained using more advanced classifiers, such as Logistic, RandomForest, and IBk classifiers.

Table 5: Machine Learning Classifiers Results

Org.	Measure	ZR	OR	J48	NB	IBK K=1	IBK K=3	IBK K=10	LG	RF
S1	Accuracy	<b>76.11</b>	68.36	71.93	72.96	65.28	74.32	75.17	73.63	67.67
	F-measure	0	0.11	0.01	<b>0.29</b>	0.25	0.27	0.15	0.08	0.24
	AUC	0.5	0.49	0.46	0.57	0.53	<b>0.64</b>	0.61	0.37	0.57
S2	Accuracy	54.78	60.45	62.2	63.03	63.33	61.34	<b>65.13</b>	60.99	58.75
	F-measure	0	0.55	0.5	0.42	0.57	0.55	<b>0.65</b>	0.48	0.55
	AUC	0.5	0.6	0.6	<b>0.66</b>	0.63	0.64	<b>0.66</b>	0.64	0.6
M1	Accuracy	50.76	65.73	66.47	63.3	61.63	67.34	65.09	<b>70.72</b>	64.67
	F-measure	0	0.63	0.59	0.47	0.6	0.64	0.64	<b>0.66</b>	0.64
	AUC	0.5	0.66	0.71	0.74	0.62	0.69	0.72	<b>0.76</b>	0.69
M2	Accuracy	85.24	85.13	85.96	82.24	79.14	82.69	86.45	<b>87</b>	83.46
	F-measure	0	0.22	0.22	0.32	0.26	0.26	0.29	<b>0.33</b>	0.30
	AUC	0.5	0.24	0.4	0.43	0.61	0.43	0.3	<b>0.7</b>	0.58
L1	Accuracy	71.4	69.15	71.61	70.79	64.4	67.36	68.43	<b>72.2</b>	66.28
	F-measure	0	0.28	0.27	0.19	<b>0.37</b>	0.34	0.33	0.11	<b>0.37</b>
	AUC	0.5	0.49	0.53	0.55	0.58	0.6	<b>0.65</b>	0.59	0.61
L2	Accuracy	50.22	49.91	52.92	53.9	57.11	58.53	58.66	<b>58.88</b>	55.71
	F-measure	0	0.48	0.38	0.23	0.57	0.58	<b>0.61</b>	0.55	0.57
	AUC	0.5	0.33	0.44	0.29	0.39	0.28	0.24	0.5	<b>0.57</b>

## 5 Communities Formed by Employees

### 5.1 Community Detection Algorithm

In order to better understand the structure of each organization we used Cytoscape’s Girvan-Newman fast greedy algorithm implementation [6] to separate each informal social network into disjointed communities. Each community is marked with a different color in Figures 1-6. Node shapes in these figures indicate whether the particular Facebook user holds a management position in the organization. Triangle nodes represent users who, to the best of our knowledge, held management positions, while

square nodes represent users who to the best of our knowledge did not hold any management position. Circles represent Facebook users' holding an unknown position within the organization.

## 5.2 Community Role Analysis

After separating the informal social network of each organization into disjoint communities, we analyzed the role of all the major organization's communities (see Table 6). We cross-referenced the community members with position descriptions and locations of residence from their Facebook profile pages. We also randomly chose several dozen users from each community. For these users, we manually inspected the user's positions within the organization by using publicly available sources, such as LinkedIn. During this process, we reviewed several thousand employees' profiles and identified the organizational positions of 4,767 users. The role of each community in the organization was determined by the majority of the community members' positions, geographic locations, and employment histories. For example, in case most of the sampled community users lived in New York City and worked as software developers in the organization, then we determined that the community is part of the organization's R&D division in New York City. By understanding the role of each community, we inferred details about the organization and the people it employed. The roles of the different communities within the targeted organizations are presented in Sections 5.2.1- 5.2.6.

### 5.2.1 S1 Communities

The community detection algorithm separated the S1 organization social network into five main communities (see Figure 1). Community role analysis revealed that S1 has several branches in Asia, most of them consisting of R&D employees. There were four R&D communities consisting of employees with different sets of skills. While three communities included mainly software developers (blue, red, and gray communities in Figure 1), one community consisted mainly of Hardware developers (yellow community). Moreover, by reviewing the users' publicly available employment history, we identified a previously acquired start-up company (orange community) and the social connections between the acquired company's employees and S1 employees.

### 5.2.2 S2 Communities

The S2 organizational social network was separated into seven communities by the clustering algorithm (see Figure 2). By reviewing the S2 employees' positions within the organization and user residence location, we discovered that S2 has two headquarters in the Middle East (blue, red, and orange communities in Figure 2) and North America (purple community). Moreover, we also discovered that the company has worldwide activities in four continents. Project managers (yellow community) are living in more than seven different major cities in the world. The S2 communities structures indicates that S2 has two headquarters that focus on R&D and worldwide operations which are managed by the different projects managers in each country.

### 5.2.3 M1 Communities

The M1 organizational social network graph was separated by the clustering algorithm into five well-connected communities (Figure 3). We discovered two of the company headquarters in North America (green community in Figure 3) and the two large R&D divisions (blue and orange communities). Moreover, we succeeded in detecting the company's Senior management community (yellow community) and the informal connection between the company's Senior managers. Identifying the Senior management position's community may assist in inferring management and key positions inside the M1 organization that in many cases were not available through publicly available resources.

### 5.2.4 M2 Communities

The M2 organizational social network graph was separated into four well-connected communities (Figure 4). Each community represented a group of R&D and Specific Domain Experts (SDE) employees



who worked in the company’s South Asia branch. Each one of the four employee’s groups was well-connected to other company employees groups which were located in different parts of the globe. For example, the South Asian yellow employees group was well-connected to other employees groups that were located in Africa, while the red employees group was well-connected to groups of employees in Australia, Europe, and North America.

### 5.2.5 L1 Communities

Using the community detection algorithm, we separated the L1 social network into 21 communities (Figure 5). Fourteen of these communities represented nine different roles inside the organization. By examining only the residence and position information of these communities, it is possible to pinpoint the group of support engineers in South America (blue community in Figure 5). We also succeeded in detecting the company’s Marketing and Sales division in Eastern Europe (yellow and purple communities ) and the cooperation’s R&D division in North America and East Asia (black community). Moreover, we discovered part the company’s R&D group in the Middle East and part of the North American Management and Sales group (green community).

### 5.2.6 L2 Communities

Using the community detection algorithm, we separated the L2 into four communities (Figure 6): two well-connected communities that contained many of the company’s R&D, consultants, and management employees in East Asia headquarters (blue and yellow communities in Figure 6). We also revealed one of the company’s amateur sport team (green community). Moreover, we were successful in detecting the cooperation’s international Senior management and their informal connection across four continents and more than 20 major cities (red community). By analyzing the cooperation’s international Senior management community, we could discover the cross Atlantic connections between the different cooperation’s branches.

## 6 Conclusions

This paper presents methods and algorithms that can be used to collect and analyze organizational social networks from publicly available data sources. In order to collect organization datasets, we presented crawling algorithms based on the homophily principle (see Algorithm 1) which can collect organizational data from online social networks like Facebook. We then used these crawling algorithms to collect data from six organizations’ social networks by collecting data from the organization’s employee Facebook profiles. In contrast to the BFS social network crawler, which inefficiently collected organization data, the presented organization social network crawler succeeded in collecting data of 17,096 users from six different organizations with an average precision rate of 22.1% (Table 1). We then used the collected organizational data to construct and analyze the informal social network between each of the organization’s employees. We then calculated eight centrality measures for each user in the organization. We used these centrality measures in order to uncover leadership roles inside the organization (Section 4). We discovered that the organizations’ users who received relatively high values in one of the centrality measures were more likely to hold management positions inside the organization. Furthermore, the closeness centrality measure presented the best precision at 20 results, with an average precision at 20 of 76% (see Table 3). Using the closeness centrality measure, we identified 85 management positions roles where 40% of the 84 positions were hidden management roles, and did not appear in the users’ Facebook profiles pages (see Table 4).

In Section 4.2, we presented a method for identifying organization management positions by using Machine-Learning algorithms. Using the WEKA software, we tested and evaluated several algorithms on the collected organization’s datasets which were based solely on the calculated centrality measures. All the evaluated classifiers returned better accuracy results than the trivial ZeroR classifier. Moreover, better AUC and F-Measure results were obtained using more advanced classifiers, such as Logistic, RandomForest, and IBk classifiers (see Table 5). We believe that these classification results can be improved by adding more features, such as the employees’ age, gender, and the organization’s seniority

to the classification algorithm. Moreover, similar Machine-Learning techniques can be applied to identify other positions inside the organization, such as developers, sales representatives, support engineers, and Senior managers.

In this study we also used the community detection algorithm in order to separate each organization's social network into disjointed communities (see Figures 1- 6). Next, by identifying the positions of more than 4,750 of the organization's employees, we discovered each community role and geographic location according to the held positions and living locations of the majority of community's users. Using this method, we succeeded in inferring many observations about each organization. For each organization, we discovered the geographic locations of its branches and the common employees' qualifications in each branch. We also discovered several non-trivial insights on each company. We discovered that although S1 acquired a start-up R&D company, the acquired company still performed as a separate company with almost no social connections to S1. This type of discovery can be used by the organization's management to identify problems in the social structure of the company, such as structural holes [4]. We discovered that S2 had many project managers who worked in different countries across the world. In M1 and L2, we uncovered the company's Senior management community and their informal friendship connections. Detecting the organization's Senior management community can assist in identifying other hidden management and key positions inside the organization. Furthermore, by understanding the connections between the company's Senior managers, we can reveal the connections between the different organization's branches. In M2, we could infer the Asian branch methods of work where each discovered group inside the Asian branch consisted of R&D and Specific Domain Experts employees who worked with company's employees in different continents. In L1, we discovered the company's support divisions in South America and the company's Sales and Marketing division in Eastern Europe.

We believe this study has several future research directions. A possible direction is to create multi-label organizational social network by cross-referencing the organization's online social network with other organization's social networks, such as the social network created by the organization's emails [34]. These a multi-label social networks can assist in better understanding the organization. Another possible direction is to combine different community detection algorithms in order to improve the organization's community detection results and discover more communities inside each organization. Another possible direction is to enrich the organization's user collected data by automatically adding user data from different publicly available data sources, such as LinkedIn and people search engines. Adding more details to the collected organization's users can improve the results of the organization's community roles identification process.

## 7 Availability

Anonymous version of the organizations' social network topologies are available for other researchers to use on our research group website <http://proj.ise.bgu.ac.il/sns/>.

## References

- [1] A. Acquisti and R. Gross. *Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook*. 2006.
- [2] T. Allen and S. Cohen. Information flow in research and development laboratories. *Administrative Science Quarterly*, pages 12–19, 1969.
- [3] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. The socialbot network: When bots socialize for fame and money. 2011.
- [4] R. Burt. *Structural holes: The social structure of competition*. Harvard Univ Pr, 1995.
- [5] C. Campbell, P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 528–531. ACM, 2003.



- [6] A. Clauset, M. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.
- [7] C. Dwyer, S. Hiltz, and K. Passerini. Trust and privacy concern within social networking sites: A comparison of facebook and myspace. In *Proceedings of AMCIS*, pages 1–12. Citeseer, 2007.
- [8] E. Estrada and J. Rodriguez-Velazquez. Subgraph centrality in complex networks. *Physical Review E*, 71(5):056103, 2005.
- [9] Facebook-Newsroom. <http://www.facebook.com>.
- [10] M. Fire, O. Tenenboim, L. Puzis, R. Lesser, L. Rokach, and Y. Elovici. Computationally efficient link prediction in variety of social networks. 2012.
- [11] L. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.
- [12] M. Gjoka, C. Butts, M. Kurant, and A. Markopoulou. Multigraph sampling of online social networks. *Selected Areas in Communications, IEEE Journal on*, 29(9):1893–1905, 2011.
- [13] A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, 2008.
- [14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18, November 2009.
- [15] E. Jacobson and S. Seashore. Communication practices in complex organizations. *Journal of Social Issues*, 7(3):28–40, 1951.
- [16] M. Kilduff and D. Brass. Organizational social network research: Core ideas and key debates. *The Academy of Management Annals*, 4(1):317–357, 2010.
- [17] M. Kilduff and W. Tsai. *Social networks and organizations*. Sage Publications Ltd, 2003.
- [18] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.
- [19] V. Krebs. Mapping networks of terrorist cells. *Connections*, 24(3):43–52, 2002.
- [20] M. McPherson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.
- [21] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC’07)*, San Diego, CA, October 2007.
- [22] M. Newman. A measure of betweenness centrality based on random walks. *Social networks*, 27(1):39–54, 2005.
- [23] M. Newman. The mathematics of networks. *The New Palgrave Encyclopedia of Economics*, 2, 2008.
- [24] M. Newman et al. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *PHYSICAL REVIEW-SERIES E-*, 64(1; PART 2):16132–16132, 2001.
- [25] Nielsen. [http://blog.nielsen.com/nielsenwire/online\\_mobile/august-2011-top-us-web-brands](http://blog.nielsen.com/nielsenwire/online_mobile/august-2011-top-us-web-brands).
- [26] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

- [27] K. Provan, A. Fish, and J. Sydow. Interorganizational networks at the network level: A review of the empirical literature on whole networks. *Journal of management*, 33(3):479–516, 2007.
- [28] D. Pugh, D. Hickson, C. Hinings, and C. Turner. Dimensions of organization structure. *Administrative science quarterly*, pages 65–105, 1968.
- [29] J. Rooksby, A. Kahn, J. Keen, I. Sommerville, and J. Rooksby. Social networking and the workplace. *The UK Large Scale Complex IT Systems Initiative*, pages 1–39, 2009.
- [30] P. Shannon, A. Markiel, O. Ozier, N. Baliga, J. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.
- [31] M. Sparrow. The application of network analysis to criminal intelligence: An assessment of the prospects. *Social networks*, 13(3):251–274, 1991.
- [32] C. Steinfield, J. DiMicco, N. Ellison, and C. Lampe. Bowling online: social networking and social capital within the organization. In *Proceedings of the fourth international conference on Communities and technologies*, pages 245–254. ACM, 2009.
- [33] N. Tichy, M. Tushman, and C. Fombrun. Social network analysis for organizations. *Academy of Management Review*, pages 507–519, 1979.
- [34] J. Tyler, D. Wilkinson, and B. Huberman. E-mail as spectroscopy: Automated discovery of community structure within organizations. *The Information Society*, 21(2):143–153, 2005.
- [35] D. Wilkinson and B. Huberman. A method for finding communities of related genes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5241, 2004.
- [36] N. Y. and M. S. Social network analysis and community mining in organizations based on email records. 2010.

Table 6: Organizations' Communities

Org.	Comm. Color	#Users	#Links	Number of Facebook Profiles with Positions	Number of Classified Users' Positions	Description
S1	Blue	30	96	6	16	R&D and administration groups in Asia
	Red	62	234	24	37	Mainly hardware verification engineers and chip designers in Asia
	Yellow	10	13	3	8	Hardware R&D
	Orange	46	197	13	21	Acquired startup company
	Gray	17	29	5	11	R&D in Asia
S2	Blue	10	16	5	6	IT group in the Middle East
	Red	109	645	25	45	R&D groups in the Middle East
	Orange	48	230	16	26	R&D groups in the Middle East
	Yellow	100	575	39	58	Mangers and international PM
	Purple	4	5	1	1	Group in North America
	Gray	4	6	1	1	European group
	Cyan	39	155	15	27	R&D teams in Australia and the Middle East
M1	Blue	467	7,685	100	163	R&D division
	Red	425	11,706	86	129	Senior management
	Orange	217	2,526	46	75	R&D divisions
	Yellow	254	3,023	47	51	International consultants and support engineers
	Green	23	95	4	7	North American Headquarter
M2	Blue	1,329	23,549	504	498	R&D and SDE connected to North American and Asian employees
	Red	1,071	16,637	437	430	R&D and SDE connected to Australia, Europe and North America
	Yellow	1,348	24,080	556	551	R&D and SDE connected to Africa, North America, and Asia
	Green	921	1,058	33	32	R&D and SDE connected to North America and Asia employees
L1	Blue	141	148	45	50	South America support engineers
	Red	1,461	1,934	471	473	South American Branch (IT, PM, Support engineers, and Analysts)
	Yellow	15	172	6	7	Eastern European Pricing Analysts
	Orange	1,613	7,407	448	422	South American Branch (Management, Sales, Marketing, PM, Support engineers, and Administration)
	Purple	443	13,837	100	110	Eastern European (Marketing, Sales and Pricing) consultants and support engineers
	Green	921	1,482	243	246	Middle East R&D and North American Headquarters (Management and Sales)
	Gray	774	17,247	201	175	European Consultants and Sales and South Asian Analysts.
	Cayn	154	151	46	67	East Asian - R&D
	Black	143	146	9	9	North American Branch, East Asian - R&D
L2	Blue	2,285	42,230	220	140	East Asian Headquarter (management and consultants)
	Red	1,573	19,841	605	449	International Senior management (Senior management, Senior researchers)
	Yellow	1,588	19,023	264	218	East Asian Headquarter (R&Ds and consultants)
	Green	78	1,478	4	1	The company's amateurs sport team